United Nations ECE/CES/2017/15/Add.4

Economic and Social Council

Distr.: General
9 June 2017

English only

Economic Commission for Europe

Conference of European Statisticians

Sixty-fifth plenary session
Geneva, 19-21 June 2017
Item 9 (a) of the provisional agenda
Programme of work of the Statistics subprogramme of the United Nations Economic Commission for Europe

Implementation of the UNECE Statistical Programme 2016

Addendum

Report of the Workshop on Implementing Standards for Statistical Modernisation

Note by the Secretariat

Summary

The Workshop on Implementing Standards for Statistical Modernisation was held on 21-23 September 2016 in Geneva.

The addendum is the report of that workshop and is provided to inform the Conference of European Statisticians on the implementation of the UNECE Statistical Programme 2016 in regards to its outcomes and recommendations.

GE.17-09469(E)
1709469

Please recycle

I. Introduction

1. The Workshop on Implementing Standards for Statistical Modernisation was held in Geneva, Switzerland, from 21–23 September 2016. Participants in this workshop comprised representatives from Australia, Belgium, Canada, Colombia, Croatia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Kyrgyzstan, Latvia, Luxembourg, Mexico, Netherlands, Norway, Poland, Republic of Korea, Romania, Slovenia, Spain, Sweden, Switzerland, Turkey, as well as representatives from the African Development Bank, Eurasian Economic Commission, European Central Bank, International Labour Organization, International Monetary Fund, Organization for Economic Cooperation and Development (OECD), and the Statistical Office of the European Commission (Eurostat).

II. Organization of the seminar

2. Ms. L. Bratanova and Mr. S. Vale, UNECE Statistical Division, opened the workshop. Mr. K. Blomqvist of Sweden chaired the workshop.

3. The workshop was divided into five sessions, each of which addressed a particular topic related to Implementing Standards for Statistical Modernisation, as follows:

(a) Implementing the Generic Statistical Business Process Model (GSBPM) and the Generic Activity Model for Statistical Organizations (GAMSO);

(b) Implementing the Generic Statistical Information Model (GSIM);

(c) Implementing the Common Statistical Production Architecture (CSPA);

(d) Developing a roadmap for statistical modernisation and a Modernisation Maturity Model; and

(e) Updates on other standards relevant for modernising official statistics.

III. Summary of the discussion and main conclusions reached at the meeting

4. Key items to be consider for future work include:

(a) Geospatial information standards;

(b) The metadata glossary (the GSBPM part);

(c) Metadata types / roles (for example revisiting the Bo Sundgren paper);

(d) Logical Information Model (LIM) variables;

(e) Quality Indicators for GSBPM concerning data derived from administrative sources;

(f) Providing explanatory material for GSBPM;

(g) Publishing and maintaining the Modernisation Roadmap;

(h) Planning the revision of GSBPM, GSIM and GAMSO;

(i) Monitoring relevant related standards; and

(j) Cooperation with SDMX (Statistical and Technical Working Groups)

5. Two project proposals to be prepared for consideration by the High-level Group for Modernisation are:

(a) Linking the Statistical Data and Metadata eXchange (SDMX) and Data Documentation Initiative (DDI) artefacts to CSPA; and

(b) CSPA data architecture.

A. Session 1: Implementing GSBPM and GAMSO

6. This topic was organized by Ms. A. Born (Canada) and Ms. J. Linnerud (Norway). It included the following presentations.

- Turkey - TurkStat's experiences on mapping non-statistical activities using GAMSO
- Eurostat - GAMSO in context
- Columbia - Modernisation of statistical production based in the analysis of the GSBPM
- Mexico - Spanish-language version of GSBPM
- Kyrgyzstan - Implementing the GSBPM and metadata system at the National Statistics Committee of the Kyrgyz Republic
- Spain - National adaptation of GSBPM and its application in the development of a process metadata model
- UNECE Quality Indicators group - Enhancing GSBPM: Quality Indicators and uses of administrative data
- Canada - The Record Linkage Project Process Model (RLPPM)
- OECD - Checklist for SDMX Design Projects using GSBPM

7. The following points were raised in the discussions:

(a) The need for statisticians and IT specialists to work together on modernisation projects.

(b) That translations of standards require maintenance, and that linguistic considerations can make it hard to describe certain ideas in other languages, or even in different countries dialect of the same language (e.g., for Spanish). – In the context of GAMSO, Capability Management is one area that is difficult to translate into other languages, although improvements to the explanatory text should help this to be more broadly understandable.

(c) Some participants requested more explanation of the "Capability Management" activity area of GAMSO. In GAMSO, the decision to develop or improve a capability is initiated within the Strategy and Leadership area, where the task is then handed over to the Capacity Management area (for example to build a new tool), before being handed over to the Corporate Support area to provide ongoing maintenance (e.g., training users, etc.).

(d) It was suggested that some terms, such as "support function" are not used consistently between GAMSO and GSBPM.

(e) There is need for revision of the overarching elements within GSBPM v5.0, in order to make it consistent with GAMSO.

(f) GAMSO v1.0 is currently in the process of review, and due consideration is being given to other standards from the HLG and outside (e.g. The Open Group

Architecture Forum (TOGAF), Eurostat Enterprise Architecture, etc.). However, GAMSO must be more easily readable than for instance TOGAF for non-IT staff in statistical organizations.

(g) Work is already underway on a glossary for statistical metadata, which should provide common descriptions of terms used in GSBPM. There is also work underway looking at variables in GSIM, and across the GSBPM.

(h) For the next phase of work on GSBPM Quality Indicators concerning administrative-derived data, there will be the opportunity to feed into this work through the next consultation process, or more directly by volunteering to join the group that will be working on this project.

(i) Regarding improvement of the descriptions of GSBPM sub-processes in the context of administrative data, participants were encourages to make use of the GSBPM discussion forum to post suggestions for improvement to these descriptions.

(j) Record linkage may require the recording of more metadata.

(k) The checklist for SDMX design projects has been implemented in the guidelines, and is hosted on the UNECE wiki site. This work demonstrates a good example of GSBPM-SDMX integration.

B. Topic (ii) Implementing the GSIM

8. This topic was organized by K. Blomqvist (Sweden). It included the following presentations:

(a) Sweden - RUT (Register Utiliser Tool) - A Practical implementation of GSIM as support in register-based research

(b) Norway - Presentation of a new GSIM-based system for managing classifications and codelists

(c) Italy - GSIM and the Sistema Unitario dei Metadati (SUM): State of application of the standard

(d) Denmark - Fulfilling user-needs and improving quality and efficiency GSIM and other standards

(e) Sweden - Statistics Sweden GSIM Case Study

9. The following points were raised in the discussion:

(a) There was discussion about systems for managing classifications, in terms of managing revisions to and variants of classifications, but also managing classifications that are common to multiple sources of data (e.g., sex of individual or non-response reasons).

(b) In some circumstances (e.g., where data is received from multiple countries) different versions of classifications (such as the International Standard Industrial Classification of All Economic Activities (ISIC)) may need to coexist. However in national systems, this situation can often be avoided, and the date considered determines which version is used.

(c) For data obtained from public registers (held outside of the statistical office), obtaining quality information can be challenging.

(d) Using large classification tree structures in a web-accessed classification management system can be costly in terms of memory usage. The systems described in the presentations used fixed rather than floating classifications.

(e) There was discussion of the advantages and disadvantages of incorporating codelists into classification management systems. The main disadvantage is the large number of codelists, some of which may only apply to a single dataset.

(f) Incorporating codelists requires decisions about how far to go in trying to document the large number of them, but also raises questions about harmonisation (e.g., codes for yes/no/don't know responses), but also about whether data providers could be forced to use certain families of codelists.

(g) Some counties have a dedicated unit that approves the creation of new codelists in an effort to minimise the number.

(h) It was suggested that the best way forward might be to start with the concepts, and to save the logic, rather than the codelists themselves.

(i) Estonia gave an example of how unit type could be variable depending on the question of interest. – GSIM is able to handle multiple unit types if a parent-child relationship is created to distinguish them. Another option is combinations of unit types and variables.

(j) More guidance on implementing GSIM was called for, especially as many organizations do not have metadata specialists to consider these questions in depth. This would also help to promote GSIM, and to make it easier to secure financial approval for implementation projects. It is anticipated that this will be addressed by the Roadmap project, which is currently underway. However implementation is more of a design issue for the implementer, than something that can be specified within a generic model.

(k) Having a set of core GSIM objects might be helpful for those embarking on GSIM implementation. There was also a suggestion for GSIM to move toward describing dimensional data sets. Whilst a particular organisation can pick out the elements they need from GSIM, classifications and variables are important aspects of GSIM for those implementing it.

(l) Those who had experience of developing a core set of GSIM objects (Denmark, Italy, Norway and Sweden) were encouraged to share their experiences via comments in the Clickable GSIM (http://www1.unece.org/stat/platform/display/GSIMclick).

(m) It was suggested that a software implementation of GSIM might encourage implementation in the same way that software applications (such as Colectica) may have boosted implementation of DDI.

(n) It was suggested that correspondence tables could be used to communicate metadata (e.g., classifications and codelists) to users. – Particularly researchers, who are a very important group of users.

(o) The importance of incorporating geospatial information within GSIM was highlighted.

C. Topic (iii) Implementing CSPA

10. This topic was organized by Mr. D. Barraclough (OECD) and Mr. A. Hamilton (Australia). It included the following presentations:

(a) Australia - Using CSPA to transform Data Acquisition;

(b) Sweden - Summary of GSIM Statistical Classifications;

(c) OECD - The Logical Information Model and variables.

11. The following points were raised in the discussions:

(a) In the case of Australia's use of CSPA for data acquisition, an integration partner was contracted, who will integrate three software products. A best of breed solution was chosen.

(b) The Australian transformation programme requires identification of benefits from projects, and savings of around AUD$10 million on data acquisition are expected, mainly due to reduction of field staff. It is possible that the service specifications can be shared from this work.

(c) The importance of using the same capability to perform a function (wherever that function is performed) is a part of an organization's target operating model.

(d) Clear guidelines were called for on the boundary between GSIM and LIM. Whilst GSIM needs to be stable, a clear set of principles is needed on LIM. One approach might be to follow the data through the GSBPM processes, and to consider what concepts are required at each stage. It may be useful to involve other parts of an organization when developing conceptual models.

(e) It was suggested that LIM and GSIM should link to or distinguish between micro- and macro-data. However, in some cases (e.g., national accounts), statistics may be derived from semi-aggregated data, rather than microdata.

(f) Consistency between LIM and DDI/SDMX was called for. More references to GSIM in the SDMX guidelines might be useful.

(g) Codelists used in SDMX and GSIM classifications are not exactly the same thing: Whilst a codelist could correspond to a variant of a classification, several variants may be needed (e.g., for national accounts). A classification database can be used to generate codelists.

(h) In GSIM, it is possible to create a value domain for a given variable.

(i) Countries should be able to adopt these standards at their own pace, in line with the aims of the Roadmap project.

12. Following the presentations, there was a panel session, moderated by Mr. D. Barraclough (OECD) and comprising Mr. D. Grofils (Eurostat), Ms. E. Holm (Sweden) and Ms. R. Piela (Finland) as panel members. The following points were raised at during the panel session:

(a) How do we move from a new requirement to the specification of a CSPA service?

i. At the moment, one has to choose between using a service in the catalogue, look for partners to share the cost of developing a CSPA service, or can build their own. There are not yet many services available in the CSPA catalogue, so maybe ways of cataloguing requirements should be examined.

(b) What is the difference between Service-Oriented Architecture (SOA) and CSPA?

i. It was pointed out that CSPA is not the same thing as a Service Oriented Architecture, and that it comprises more than just common interfaces. Key

components include standard structures for business functions and information objects involved in statistical services.

(c) How can we be sure that CSPA services do what they say on the tin, and trust that they are of a sufficient quality?

i. Candidate services are reviewed by the CSPA group, and documentation is required for the CSPA service, in order to better understand what that CSPA service is doing. This helps to ensure that the service is not a duplication of one which already exists in the catalogue, as well as helping to understand its level of granularity;

ii. Information is provided to assist with implementation, including information on LIM. Since LIM is developed as needed, the service vendor should provide support to the service;

iii. Whilst the specification should be related to GSIM, service developers might find this hard to understand;

iv. The term "CSPA compliant" was describes as ensuring that terminology/vocabulary does not contradict other standards (e.g., GSIM) which are overseen by the High-Level Group for the Modernisation of Official Statistics (HLG-MOS), and the standardisation of its inputs and outputs, as well as its "Plug and Play" nature. Utility services related to a specific area of corporate support would not normally be considered to be a CSPA service.

(d) How do we increase the number of organizations which implement CSPA?

i. Support from Governments for CSPA would be very useful, but on the other hand if there is a sizable modernisation community, this community can support its implementation: It is then a question of whether people decide to implement it or not;

ii. One selling point is that for decentralised national statistical systems, CSPA offers benefits even within that country. Similarly, if you have implemented GSIM, then a moderate amount of extra effort will bring the benefits associated with CSPA;

iii. It is important to take a lifecycle view of the advantages and disadvantages of CSPA within your organization;

iv. If you are sharing a CSPA component, you will have an extra responsibility to maintain it because you know that other people are then relying upon it. It is important to vi

v. Early adopters may get less benefit than those who adopt it later.

(e) Say that we have a business requirement for a new service: How should one decide whether to reuse an existing service, collaborate with someone building a new service, or develop a new service entirely?

i. The first step is to consult methodologists on the soundness of the methodology, and to consult IT staff about its suitability (e.g., is it built using obsolete technology?);

ii. The crucial deciding factor on adoption is that key statistical outputs (such as national accounts) cannot be delivered late, and that responsibilities are clear in the event that a delay occurs;

iii. Another key consideration is whether senior managers are willing to pay the costs of building a new CSPA service.

(f) How can our enterprise architectures evolve to incorporate CSPA services?

i. Mapping services to functions in business processes should offer new opportunities to incorporate CSPA services in enterprise architectures;

ii. At the last CSPA hackathon, there was work on orchestration, to see how a service could be plugged into an enterprise architecture.

(g) What are the plans for further developing the CSPA standard (to maintain models, or to make it easier to use)?

i. A key element will be work on LIM to facilitate CSPA implementation. LIM is a more specified version of GSIM;

ii. More documentation and good demonstrations of particular CSPA implementations would be useful, as well as a template for developers;

iii. It was suggested that another CSPA hackathon might be useful;

iv. Trust is important for developing shared services;

v. Eurostat is already in the process of developing shared services;

vi. It terms of the relationship between CSPA and SDMX/DDI, CSPA is more a set of standard interfaces for development of services (e.g., attribute levels and polymorphism), rather than an implementation standard itself. However, CSPA does recommend SDMX as a data transfer protocol. The compatibility of CSPA with SDMX/DDI can be established by mapping GSIM to DDI and SDMX.

(h) What are the main obstacles to CSPA implementation, and how to overcome them?

i. The need for more accessible information, to raise awareness of CSPA. Simple demonstrations of creating a CSPA service (such as wrapping an R package in CSPA) would be useful;

ii. The length of time needed to produce a CSPA service is a challenge;

iii. CSPA is competing with existing marketplaces for algorithms. – These have an advantage in that you don't have to do version management. Top managements often prefer the guarantee offered by commercial services;

iv. It was noted that the CSPA repository on the European Commission website is currently hidden behind a login screen, which might deter some would-be users of CSPA;

v. Including the class library templates and service implementation description should make CSPA services easier for developers to use in their organizations;

vi. There are already some success stories, such as the CSPA service developed to transform SDMX messages (used to make correspondence to local production systems);

(i) How can we measure the successfulness of CSPA?

i. Initially, a simple test of successfulness is simply that two organizations are using a CSPA service created by a third organization;

ii. Other measures of success that were suggested included that CSPA service providers can fix any bugs within a day, or the number of participants at CSPA meetings.

D. Topic (iv) Developing a roadmap for statistical modernisation and a Modernisation Maturity Model.

13. This topic was organized by Ms. A. Born (Canada) and Ms. J. Linnerud (Norway). It included the following presentation and was followed by a groupwork exercise to test the Modernisation Maturity Model (MMM). During the discussion that followed the testing of the MMM, the following points were raised.

1. Feedback from the GAMSO group

(a) Not all organizations are aligned to GAMSO, and there are a number of country-specific adaptations of this standard. The interpretation (in the context of GAMSO) of some dimensions is easier than others, and the most important dimension for assessing the maturity of GAMSO implementation should be the "Business" dimension;

(b) The relation of GAMSO to GSBPM should be made more explicit.

2. Feedback from the GSBPM group

(a) Some participants felt that the descriptions required elaboration, and used IT-oriented language. Some in this group had difficulty with the descriptions of some of the dimensions (e.g., Applications, Information);

(b) The version of GSBPM that maturity is being assessed against should be taken into account.

3. Feedback from the GSIM group

(a) Some slight refinement of the text for consistency would be useful;

(b) There was difficulty in interpreting some of the dimensions (e.g., Business, Methods) in the context of GSIM;

(c) It is also difficult to consider GSIM in isolation with other standards;

(d) Some participants felt unable to respond to some cells in the assessment matrix (some combinations of level and dimensions);

(e) More explanation was called for.

14. The final morning of the workshop began with a panel on the Roadmap to modernisation, moderated by Mr. S. Vale (UNECE), and composed of Ms. E. Kaukonen (Finland), Ms. M. Signore (Italy), and Ms. T. Šijanec (Slovenia) as panel members. Each panel member made a presentation on the modernisation story within their organization, followed by discussions. The following points were made in these discussions:

(a) The current Roadmap project is underway, and has delivered a Modernisation Maturity Model (MMM), which is used for benchmarking an organization's maturity level. The next phase of work on the Roadmap project will identify

instruments that can help statistical organizations to raise their level of maturity from the initial level, to a higher level, in a given standard (GAMSO, GSBPM, GSIM, CSPA).

(b) A mapping of an organization's activities to GAMSO can be one way of understanding an organization's maturity level with respect to this standard. Some organizations then create code lists for each activity identified through the mapping process.

(c) The same is true for GSBPM, since both GAMSO and GSBPM are abstract models which one might map to a given national situation. Part of the benefit of these standards is that the mapping process helps to identify areas that could be rationalised.

(d) GAMSO can also be used for modelling cost allocation, and for mapping quality indicators.

(e) For quality indicators, version 1.0 of the GSBPM Quality Indicators has been released, which concerns data derived from surveys. The next phase of work will consider indicators for data obtained from administrative sources.

(f) Some participants (Finland, Slovenia) felt that revising GSBPM was not as important as work on other standards, other than work on the overlap with GAMSO.

(g) It was explained that there were plans to review GAMSO, GSBPM and GSIM as a block in 2018, but only if there was enough agreement to do this, plus a strong business case.

(h) For the time being, those wishing to suggest changes to the textual descriptions of GSBPM sub-processes could post comments in the GSBPM discussion forum.

(i) In terms of suggestions for possible outputs of the Roadmap project, to assist organizations at different stages on the journey to modernisation, the following suggestions were made:

i. Further opportunities for exchange of experiences and training;

ii. Identification of common challenges;

iii. A checklist of benefits realised;

iv. Short online videos; and

v. More promotion of the online case studies.

(a) It was emphasised that the tools outputted from the Roadmap project should allow organizations to take what they need to allow them to modernise at their own pace. They should be available on the web;

(b) It was also pointed out that a lot of good material is already available on the UNECE wiki pages, such as Clickable GSIM. One suggestion made was to have more outreach to non-UNECE countries;

(c) The session organizer thanked the participants for their feedback from the group work to test the MMM. – This feedback will be considered in future work on the Roadmap.

E. Topic (v) Updates on other standards relevant for modernising official statistics

15. This topic was organized by Mr Juan Múñoz López (Mexico), and took the format of two separate panel sessions. The first panel was composed of Mr Guillaume Duffes (France) and Ms Marie Haldorsen (Sweden), who made the following presentations:

(a) The DDI Questionnaire Generator (France); and

(b) Connections between Geospatial data and Standards (Sweden).

16. The following points were raised in the discussion:

(a) Eurostat's Vision 2020 projects attempt to integrate various different standards, including work on statistical units. Geospatial information should be connected to this work;

(b) The GEOSTAT 2 projects aims to propose a European version of a statistical framework for geospatial data. However, more interaction is needed with statistical standards, and greater collaboration with the United Nations Committee of Experts on Global Geospatial Information Management (UN-GGIM);

(c) It may be fruitful to link geospatial information to GSIM, as well as GSBPM;

(d) Regarding the DDI Questionnaire Generator, this requires DDI Lifecycle v3.2 and above as a prerequisite. If your organization used DDI Codebook, Colectica can edit this;

(e) It is desirable to be able to integrate categories, concepts and codelists in the metadata supplied to data users, for example in Resource Description Framework (RDF) form.

17. The second panel was composed of Mr. David Barraclough (OECD), Mr. Guillaume Duffes (France) and Mr. Marco Pellegrino (Eurostat), who made the following presentations:

(a) Recent developments in SDMX (OECD);

(b) Validation and Transformation Language (VTL) and the Statistical Data Catalogue standard for open data portals (StatDCAT) (Eurostat); and

(c) Linked Open Metadata (France).

18. The following points were raised in the discussion

(a) Recent developments on SDMX included:

(b) Proposals for Global Metadata Structure Definitions, to help map between local and global SDMX structures;

(c) Clickable SDMX, to allow navigation between concepts;

(d) Checklists for SDMX design projects, and for data reporters;

(e) A methodology for versioning of SDMX artefacts; and

(f) Guidelines for reporting on a non-annual cycle, time transformations, and confidentiality and embargo.

(g) StatDCAT is a specialised version of DCAT for statistics. Further coordination of this work is needed with a range of statistical and non-statistical stakeholders at both the country level and international level.

(h) For Linked Open Metadata, it was suggested that there could be synergies with work on StatDCAT and Quality Indicators for GSBPM. Further work is needed on using the Single Integrated Metadata Structure (SIMS) for deriving a data quality vocabulary for StatDCAT. Further work needs to be done on the draft RDF data quality vocabulary. It was suggested incorporating the use of RDF with the GSBPM.

(i) Work that is ongoing on LIM will consider VTL reference materials.

(j) SIMS is using RDF, so offers opportunities for quality reporting, but it is undecided whether to use a flat structure or more complex structure for this. The GSBPM Quality Indicators group might benefit from consulting Eurostat on this question.

(k) It might also be worth reviewing the proposals on Global Metadata Structure Definitions in light of developments on SIMS and StatDCAT.

(l) Work on these standards is adapting to new technologies in a number of ways, such as adoption of RDF for dissemination. Linking to the semantic web/linked open data movements is a good way to connect to much larger communities.

(m) SDMX is establishing links to other standards, such as CSPA and GSIM.

19. The chair thanked participants and organizers. The UNECE Secretariat thanked the chair and the organizers for their contribution to the workshop.